

--	--	--	--	--	--	--	--	--	--

# MULTIMEDIA UNIVERSITY

## FINAL EXAMINATION

TRIMESTER 1, 2017/2018

### TPA 7021 – DATA PREPROCESSING AND ANALYSIS

(All sections / Groups)

5 OCTOBER 2017  
10.00 a.m – 12.00 p.m  
(2 Hours)

---

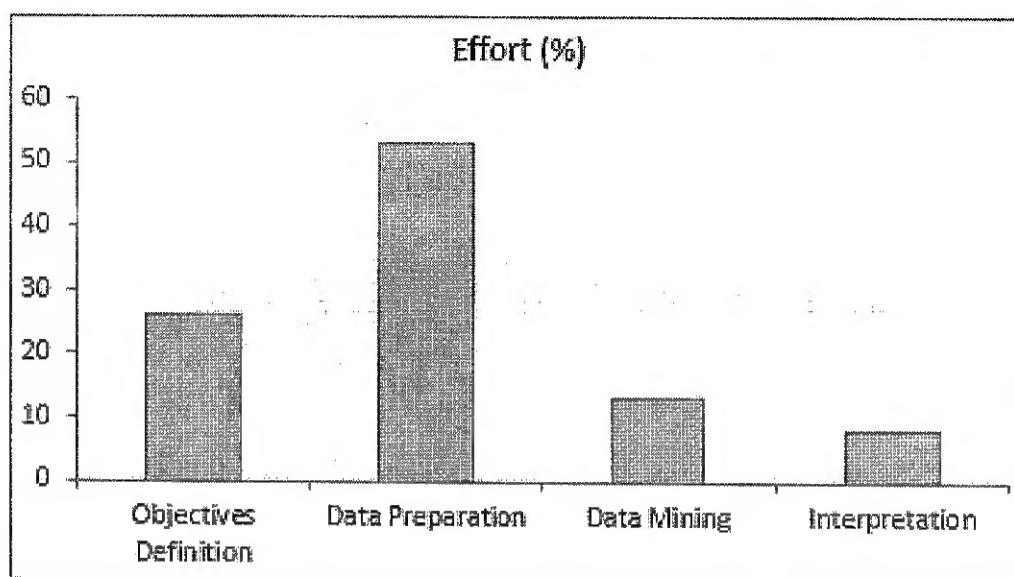
#### INSTRUCTIONS TO STUDENTS

1. This question paper consists of 6 pages with FOUR questions only.
2. Answer ALL questions.
3. Please write all your answers in the Answer Booklet provided.

**QUESTION 1**

- (a) Study the following chart and discuss the reasons for the large effort in “Data Preparation” phase.

[2 marks]



- (b) Real-world data is often “dirty”. State the characteristics of “dirty” data.

[3 marks]

- (c) Define *outlier* in the context of data preprocessing. Discuss whether all outlier should be treated as dirty data. Give an example to support your argument.

[5 marks]

Continued...

**QUESTION 2**

- (a) What chart should one use to illustrate the retail store visit patterns in a shopping mall? Assuming that the visit patterns are always *linear*.

[1 mark]

- (b) Differentiate *Nominal* and *Ordinal* types of attributes.

[4 marks]

- (c) The `ifelse()` function in R programming can act as a “flag” to data that has violated the constraints. Study the data frame `dt.Results` below carefully and write an R script to set `Marks > 80` to “A”, otherwise “B”. Create a new variable named `chkResults` to store the results.

**dt.Results**

Student_Name	Marks
Ting	75
Chong	86
Fatimah	90

[2 marks]

- (d) Which type of chart is best for visualizing the relationship between two *numerical* variables. Discuss your reason for using that chart.

[3 marks]

Continued...

**QUESTION 3**

- (a) The scenario below is an experiment gathered to investigate the performance of three different brands of laptop. Study the following R codes carefully.

```
>Laptop1 <- c(2,3,7,2,6)
>Laptop2 <- c(10,8,7,5,10)
>Laptop3 <- c(10,13,14,13,15)
>Combined_Groups <- data.frame(cbind(Laptop1,
                                       Laptop2,Laptop3))
>Stacked_Groups <- stack(Combined_Groups)
>Anova_Results <- aov(values ~ ind, data = Stacked_Groups)
>TukeyHSD(Anova_Results)
```

```
Tukey multiple comparisons of means
 95% family-wise confidence level
```

```
Fit: aov(formula = values ~ ind, data = Stacked_Groups)
```

```
$ind
      diff      lwr      upr      p adj
Laptop2-Laptop1  4 0.4206853  7.579315 0.0286585
Laptop3-Laptop1  9 5.4206853 12.579315 0.0000598
Laptop3-Laptop2  5 1.4206853  8.579315 0.0075279
```

As data scientist, what can you conclude?

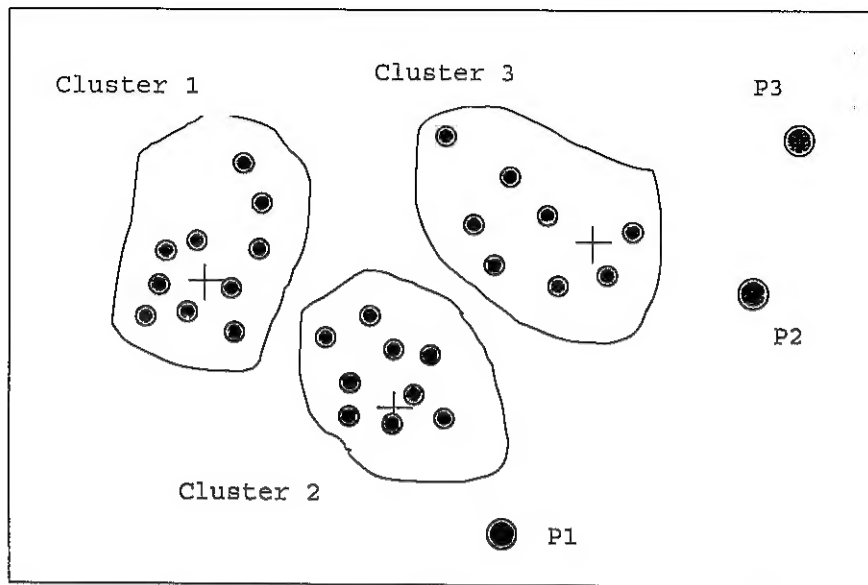
[2 marks]

- (b) What is *Noisy* data? List FOUR techniques to handle noisy data.

[3 marks]

Continued...

(c) What is the best way to handle P1, P2, and P3 if you cannot omit them?



[2 marks]

(d) What are the THREE different techniques for dimension reduction?

[3 marks]

Continued...

**QUESTION 4**

- (a) Interquartile range of an observation variable is the difference between its upper and lower quartiles. What information can you observe from interquartile range?

[2 marks]

- (b) Study the following result:

```
> TukeyHSD(Anova_Results)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = values ~ ind, data = Stacked_Groups)

$ind
      diff      lwr      upr    p adj
Group2-Group1  4 0.4206853  7.579315 0.0286585
Group3-Group1  9 5.4206853 12.579315 0.0000598
Group3-Group2  5 1.4206853  8.579315 0.0075279
```

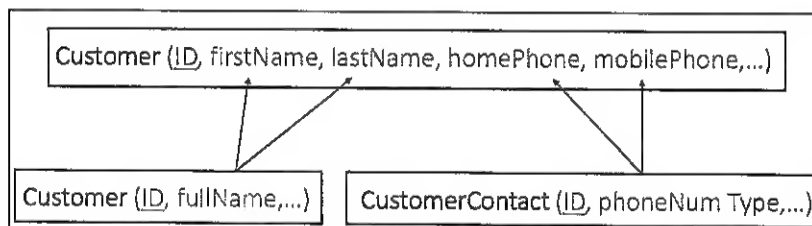
- (i) What is the command `TukeyHSD(Anova_Results)` for?

[2 marks]

- (ii) What is the conclusion based on the above result?

[2 marks]

- (c) Data integration is a challenging task. Below is an example of data integration challenge related to "Schema Integration".



Referring to figure above, how could the challenge be solved?

[2marks]

- (d) Study the R code below:

```
> Model <- lm(z ~ a + b + c + d, dataset)
> step(Model)
```

Explain how the codes above help in dimension reduction.

[2 marks]

**End of Pages.**